On the Prediction of At-Risk Patient Profiles with Big Prescription Data

Pierre Genevès[1]
pierre.geneves@cnrs.fr

Joint work with:

Thomas Calmant[1], Nabil Layaïda[1]
Marion Lepelley[2], Svetlana Artemova[2], Jean-Luc Bosson[2]

March 20, 2018

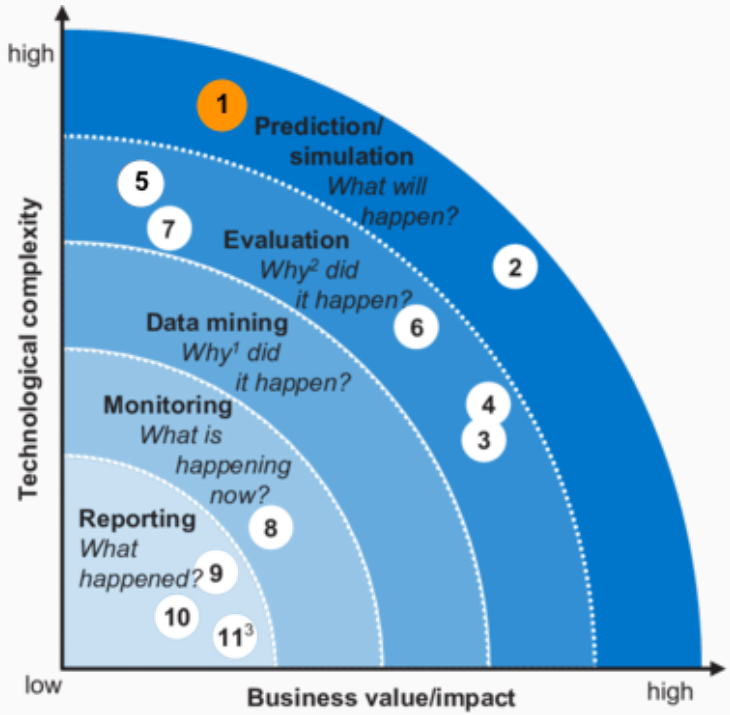
[1] Tyrex team, CNRS LIG, Inria, Univ. Grenoble Alpes, Grenoble INP, http://tyrex.inria.fr

[2] Univ. Grenoble Alpes, CNRS, Public Health department CHU Grenoble Alpes, Grenoble INP, TIMC-IMAG

A Journey in Big Data for Healthcare

Diverse data sources: sensors, prescription, genome, billing, clinical...

A spectrum of promising big data applications:



1 Risk stratification/patient identification for integrated-care programs
2 Risk-adjusted benchmark/simulation of hospital productivity
3 Identification of patients with negative drug–drug interactions
4 Identification of patients with potential diseases ("patient finder")
5 Evaluation of clinical pathways
6 Evaluation of drug efficacy based on real-world data
7 Performance evaluation of integrated-care programs and contracts
8 Identification of inappropriate medication
9 Systematic reporting of misuse of drugs
10 Systematic identification of obsolete-drug usage
11 Personal health records

[1] Machine based: evaluation of data correlations only.
[2] Hypothesis based: integration of advanced analytics to determine causation, interdependencies.
[3] Higher business value expected if further enhanced and rolled out as personal health record.

What are the associated **challenges**? Rather Math/Stat/Comp. Sci.?

→ We proposed a predictive analytics use case and addressed it.

Prediction of Adverse Effects

Adverse Effects

Undesired harmful effects resulting from medical care

e.g. hospital-acquired infection (HAI), admission in intensive care unit (ICU), pressure ulcers (PU), death

Prediction

- Almost half of adverse effects "*clearly or likely preventable*"[1]
- Crucial (hard) requirement: precise identification of at-risk profiles
  - with adapted prevention: some adverse effects could be avoided
  - ex: risk of ICU admission? $\rightarrow$ better room placement/surveillance

[1] Physicians: see [Levinson 2010], US Department of Health and Human Services

Problem

Can we predict, on the day of hospital admission, future occurrence of adverse effects?

State-of-the-Art and Open Questions

State-of-the-Art

- Traditional approach: use a scalar aggregate (score) computed from electronic health records
  - Medication Regimen Complexity Index (MRCI) [Georges et al. 2004]
  - Higher levels of MRCI at admission known to be correlated with higher risks of occurence of complications [Lepelley et al. 2017]

Open Questions:

1. What if we use all high-resolution data available to build predictors? (instead of designing aggregates) Is it feasible? Scalable? How fast?
2. Can we measure the effect of volume and variety of big drug prescription data?
3. What are the computing bottlenecks in practice?

Outline

1. The system we developed, using distributed machine learning
2. Experimental results with data of millions of patients
3. Elements of answers to the previous questions

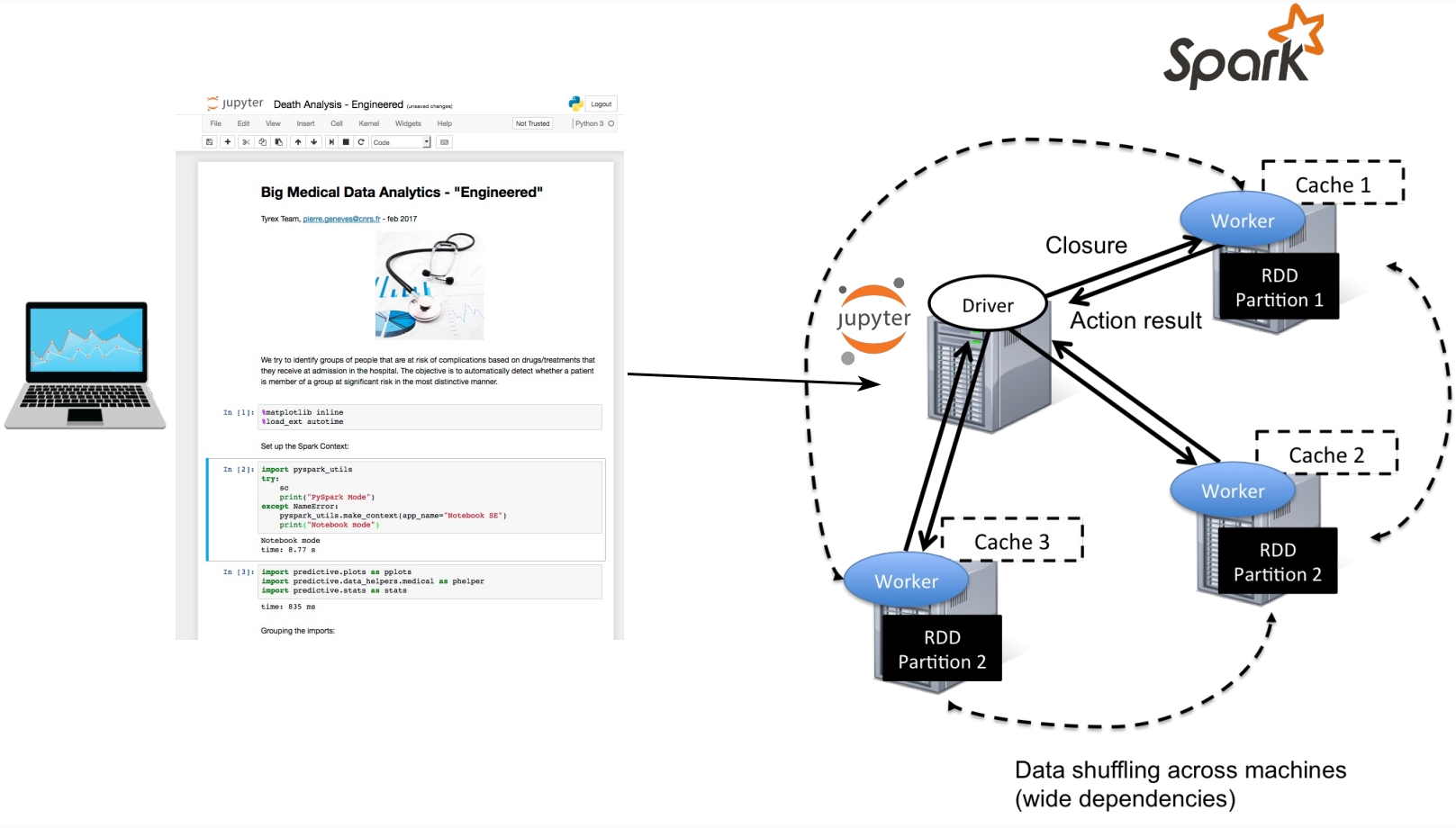
Key Methodological Aspects

Initial Postulate: drug prescription data on the day of admission contain rich information about the patient's situation and perspectives of evolution.

No prior clinical knowledge for the definition of features.
$\rightarrow$ We use supervised learning to extract this information

Evaluation of model quality and performance metrics
$\rightarrow$ k-fold cross-validation, ROC and PR AUC on imbalanced test sets, running times for distributed execution.

System Overview



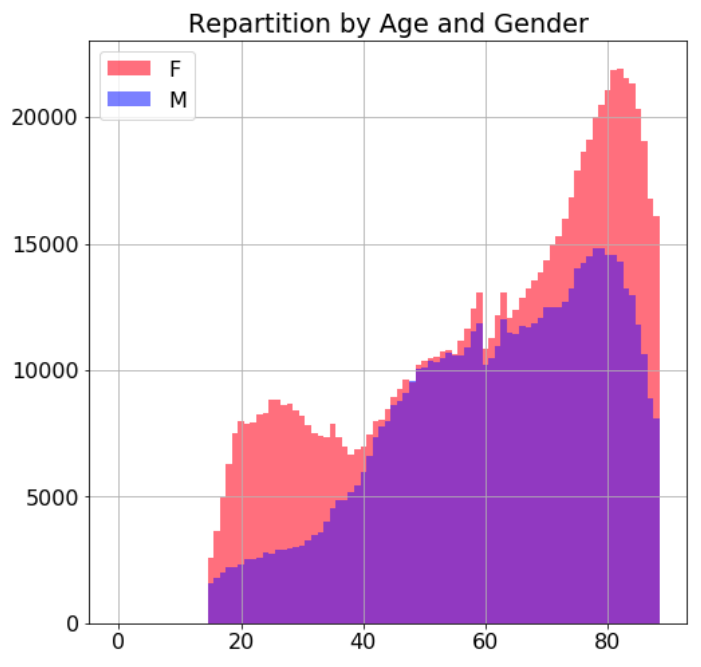
Distributed implementations of supervised ML algorithms to ensure scalability of model construction.

→ Key Technologies used: Spark, Spark SQL, MLlib/Spark ML, Docker, Jupyter Notebooks, Python, Scala.

Considered Data

Premier Perspective Database, 2006

- 417 hospitals (USA)
- 33 million admissions
- $>$ 3 billion patient billing records (operations, drugs, everything that can be billed!... $\rightarrow$ USA!)

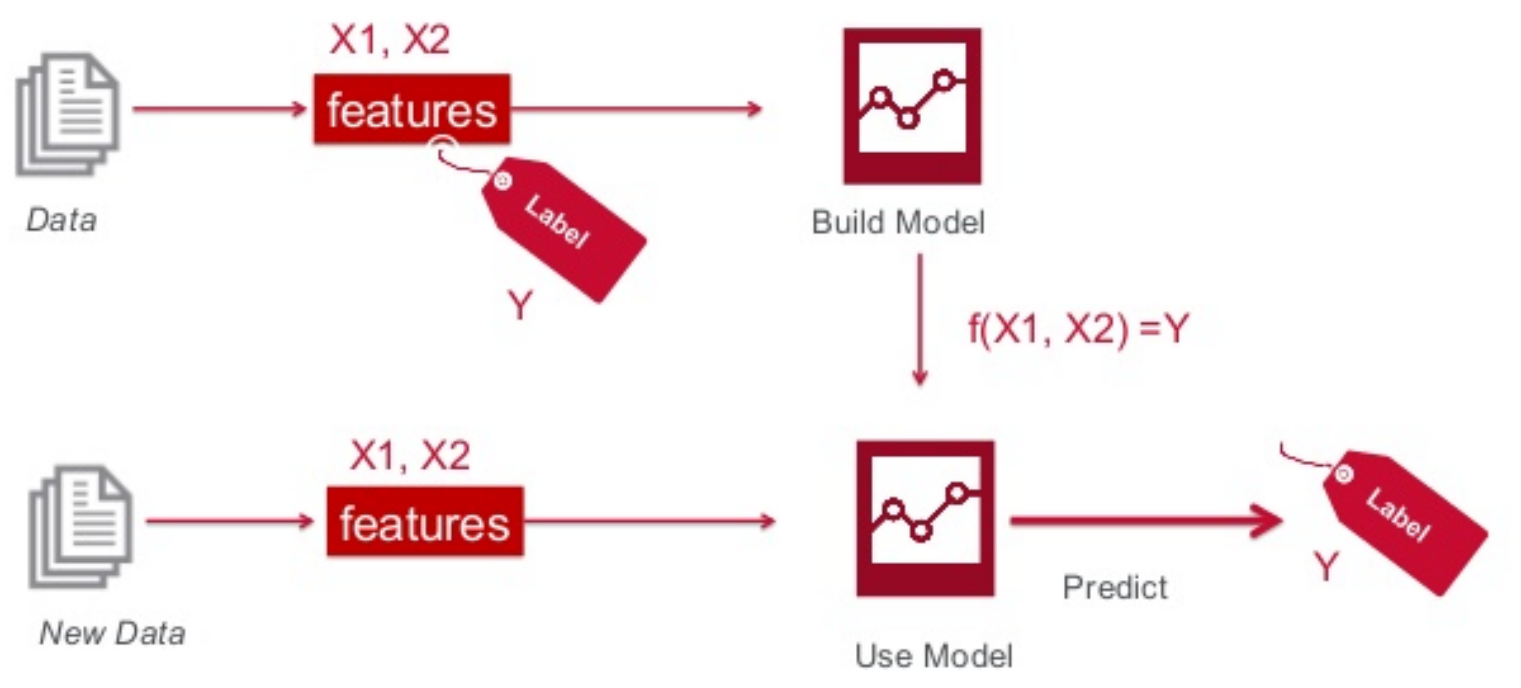


Big Data?

Our experience with the cost of initial data preparation (big joins):

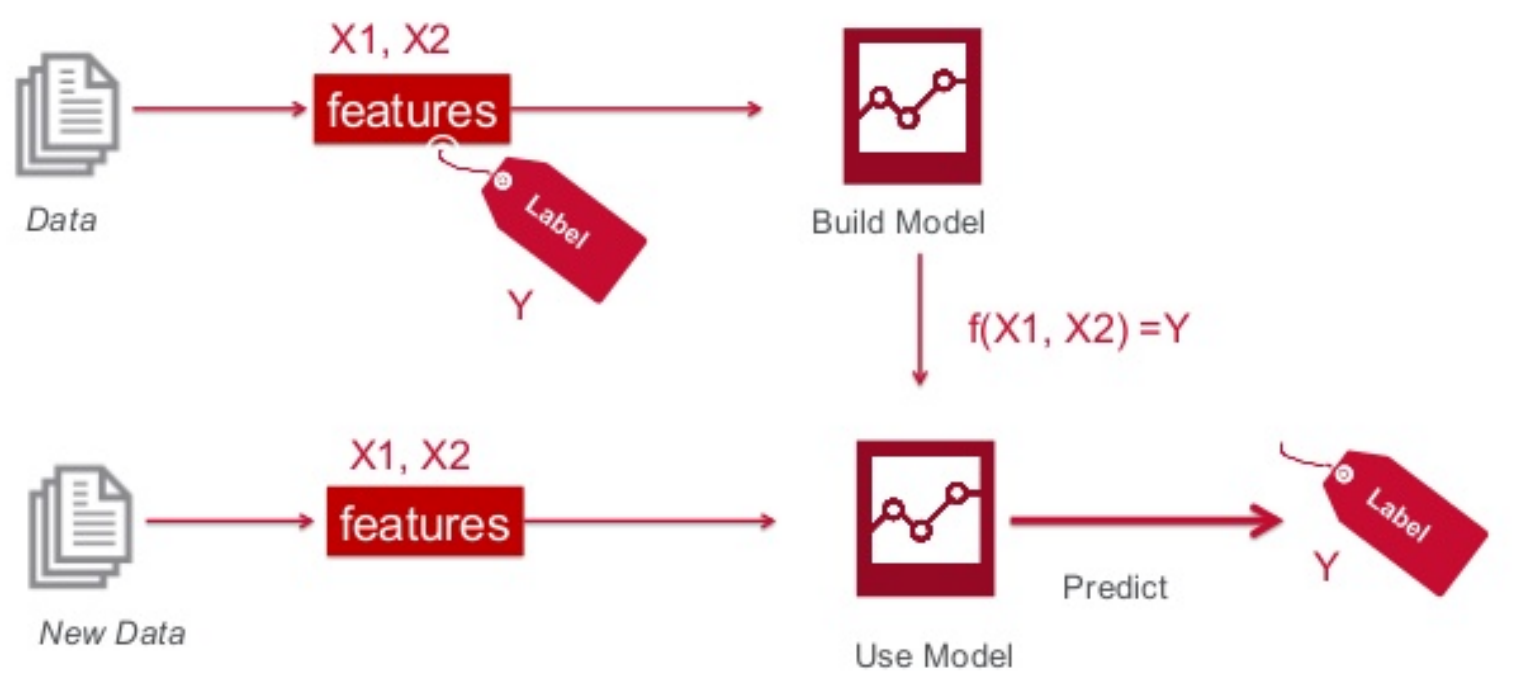
1. big join computed (by chunks) in a centralized manner: 6 days
2. distributed computations: 2 min

Distributed Supervised Machine Learning



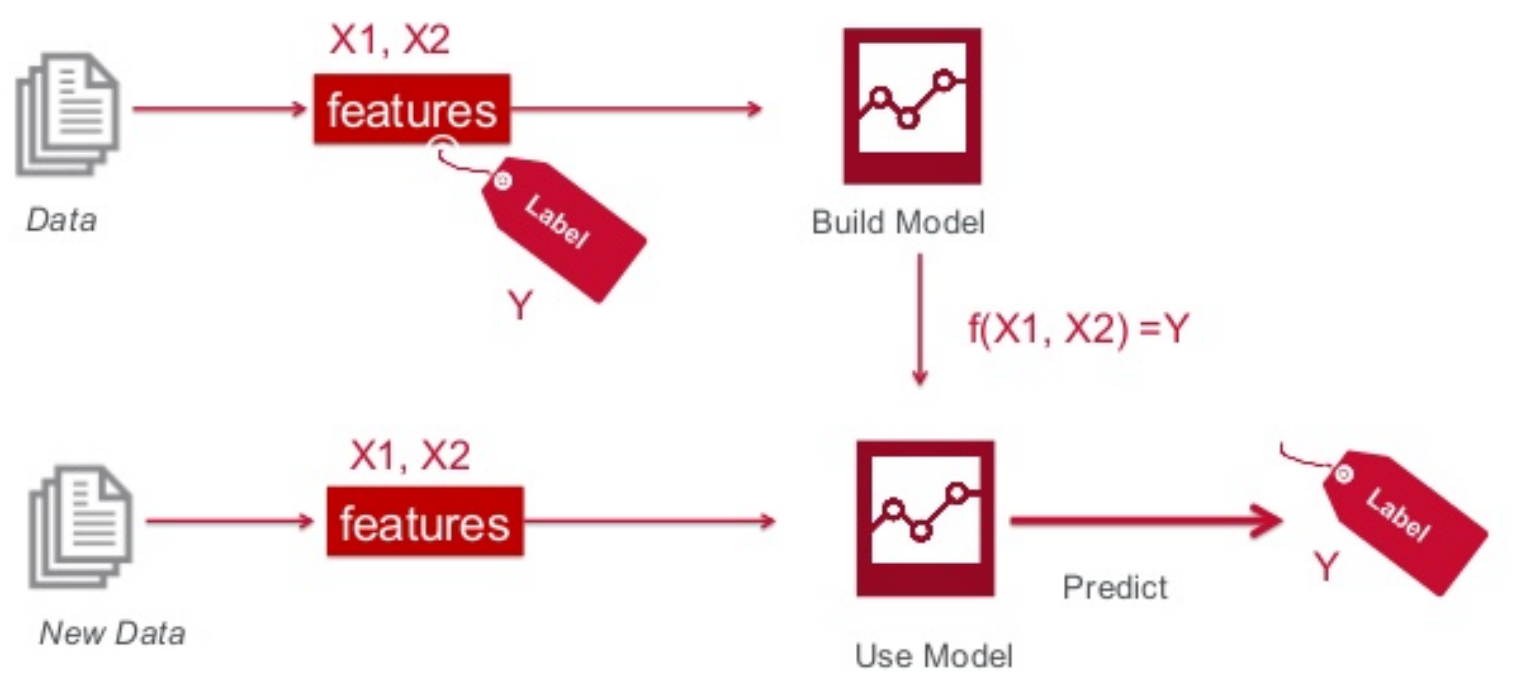
Data: billing records filtered and joined with patient info

Distributed Supervised Machine Learning



Features (`X1,X2,...XN`): extracted from patient info (age, gender...) and drug prescription data on the day of admission.

Distributed Supervised Machine Learning

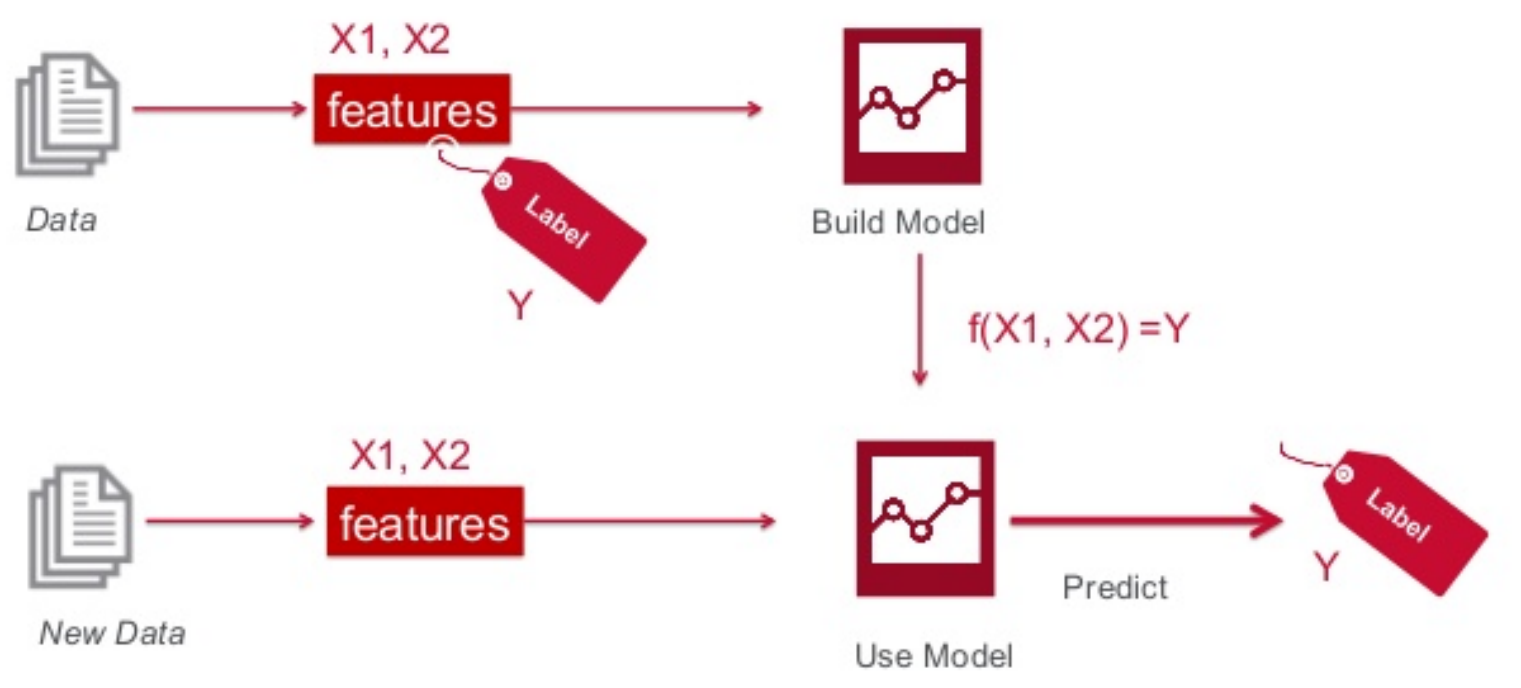


Label (Y): a boolean for each considered adverse effect (AE)

1: AE occurred

0: AE did not occur

Distributed Supervised Machine Learning



Models (f) for binary classification: Decision Trees, Random Forests, SVM, Logistic Regression, Deep Neural Networks...

Zoom on Considered Features

We group features into categories:

- B: A list of basic patient features: age, gender, admission type
- M: An aggregate score that corresponds to MRCI at admission (for comparisons)
- P: The list of drug categories served on the first day
- C: The list of detailed drug names served on the first day (with distinct variants)

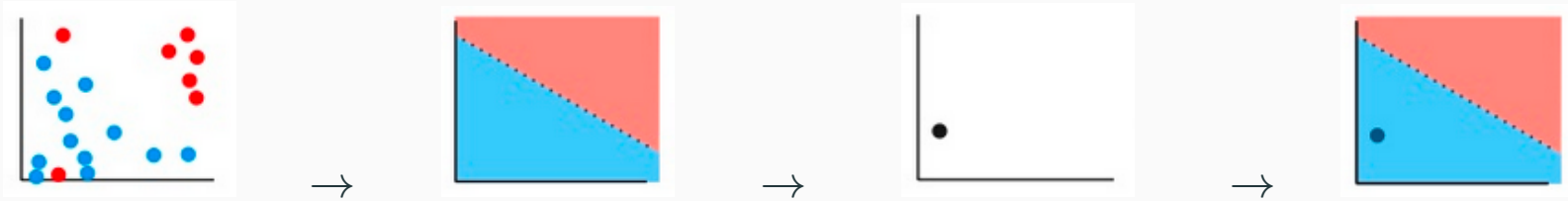
Increasing Variety/Granularity

We build models that combine feature categories: **BM**, **BMP**, **BMC**

Number of features can be large:
`length(P) > 2 000`, `length(C) > 10 000`

Binary Classification in Large Dimensions

Simple Academic Example with 2D Feature Vectors:



training data → trained model → new data point → prediction

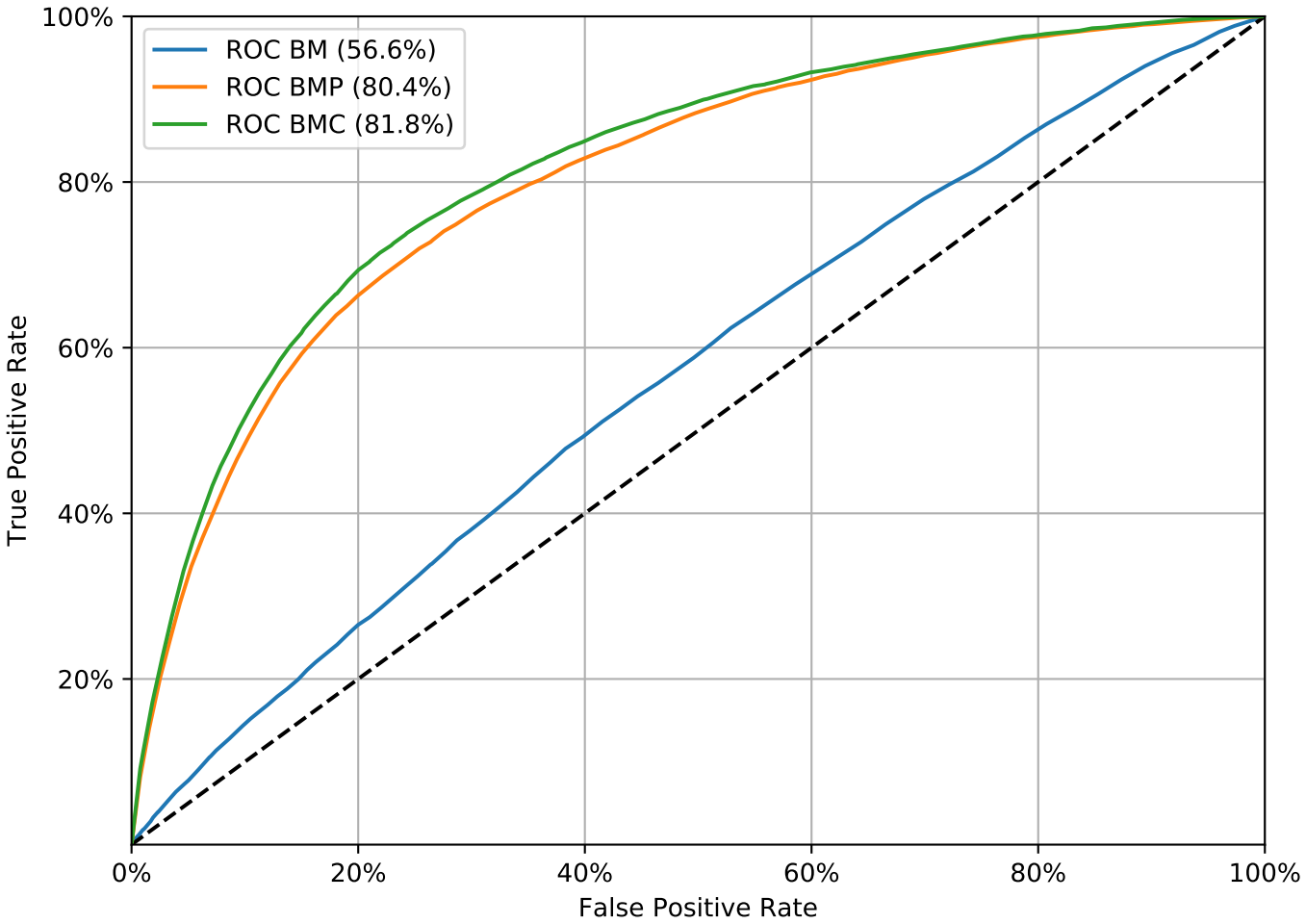
Same principle in > 10 000 dimensions

Sample feature vector in a **BMC** model

Feature Index	Feature Description	Feature Value	Standard Charge Master Code
0	Age	15	
1	Gender	1	
2	MRCI	24	
8024	DEXTROSE/NACL SOLUTION 1000ML	1.00	250258000970000
7955	NACL SOLUTION 100ML	2.50	250258000220000
7949	NACL SOLUTION 1000ML	1.00	250258000160000
7084	DOCUSATE NA CAP 100MG	1.00	250257020020000
6654	ACETAMIN TAB 325MG (EA)	2.00	250257000530000
4869	SOD BICARB INJ 8.4% 50MEQ 50ML	1.00	250250058740000
4332	POT CHL VL 20MEQ 10ML	0.50	250250053100000
3566	MORPHINE TAB SR 30MG	0.50	250250044450000
....	....	...	...

First Results: Impact of Variety

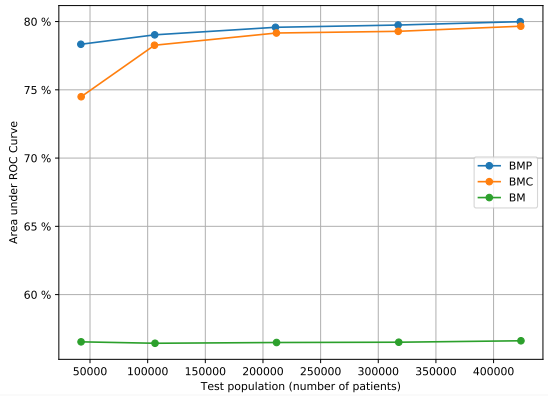
Area under the ROC curve when Predicting HAI (LR)



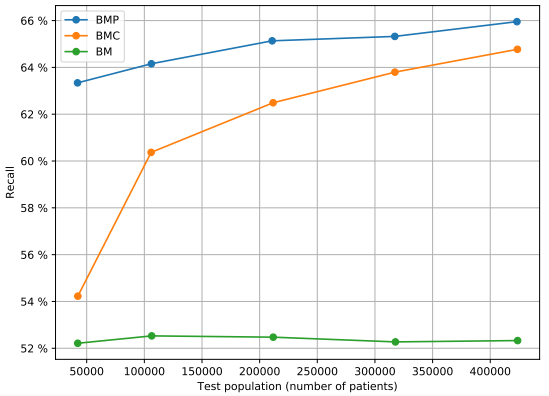
Impact of Volume (HAI Prediction)

1. Train and test sets of varying size (but with constant 2:1 ratio):

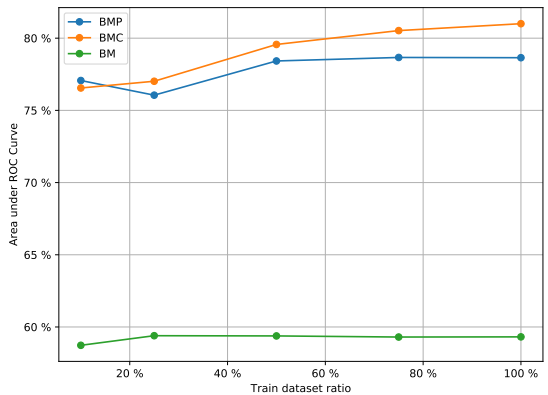
ROC AUC

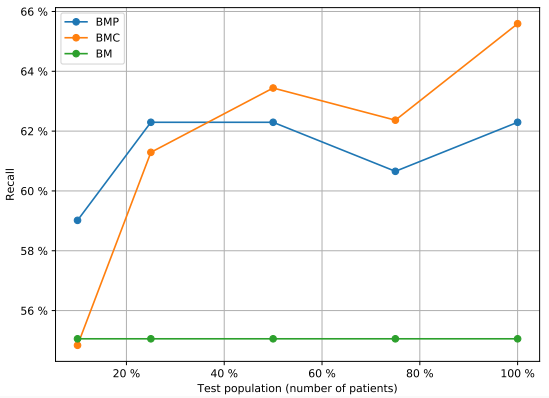


Recall (TP/P)



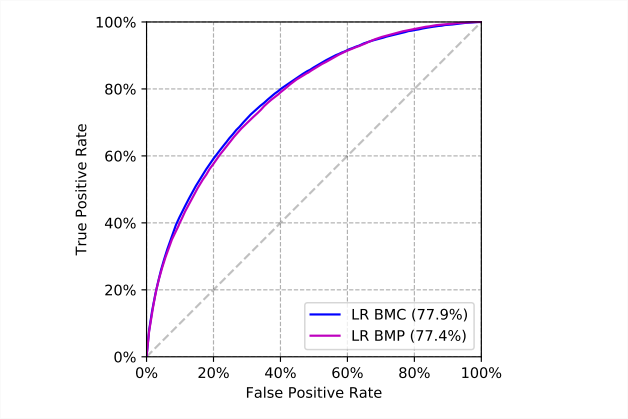
2. Train sets of varying sizes, same test set:



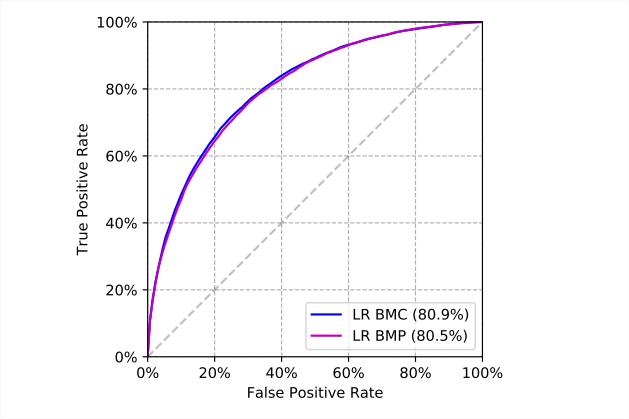


→ Increased volume tends to yield greater AuROC and greater recall.

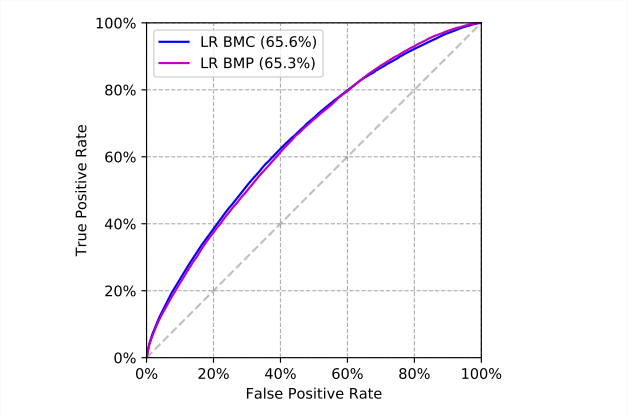
Different AuROC for Various Adverse Effects



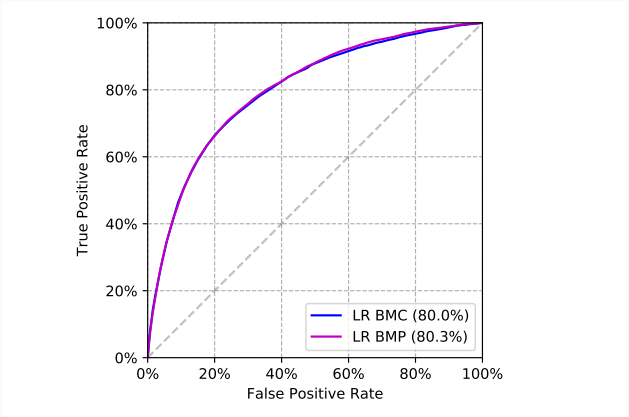
(a) Death (AUC $\geq$ 77.9%).



(b) Pressure Ulcers (AUC $\geq$ 80.9%).



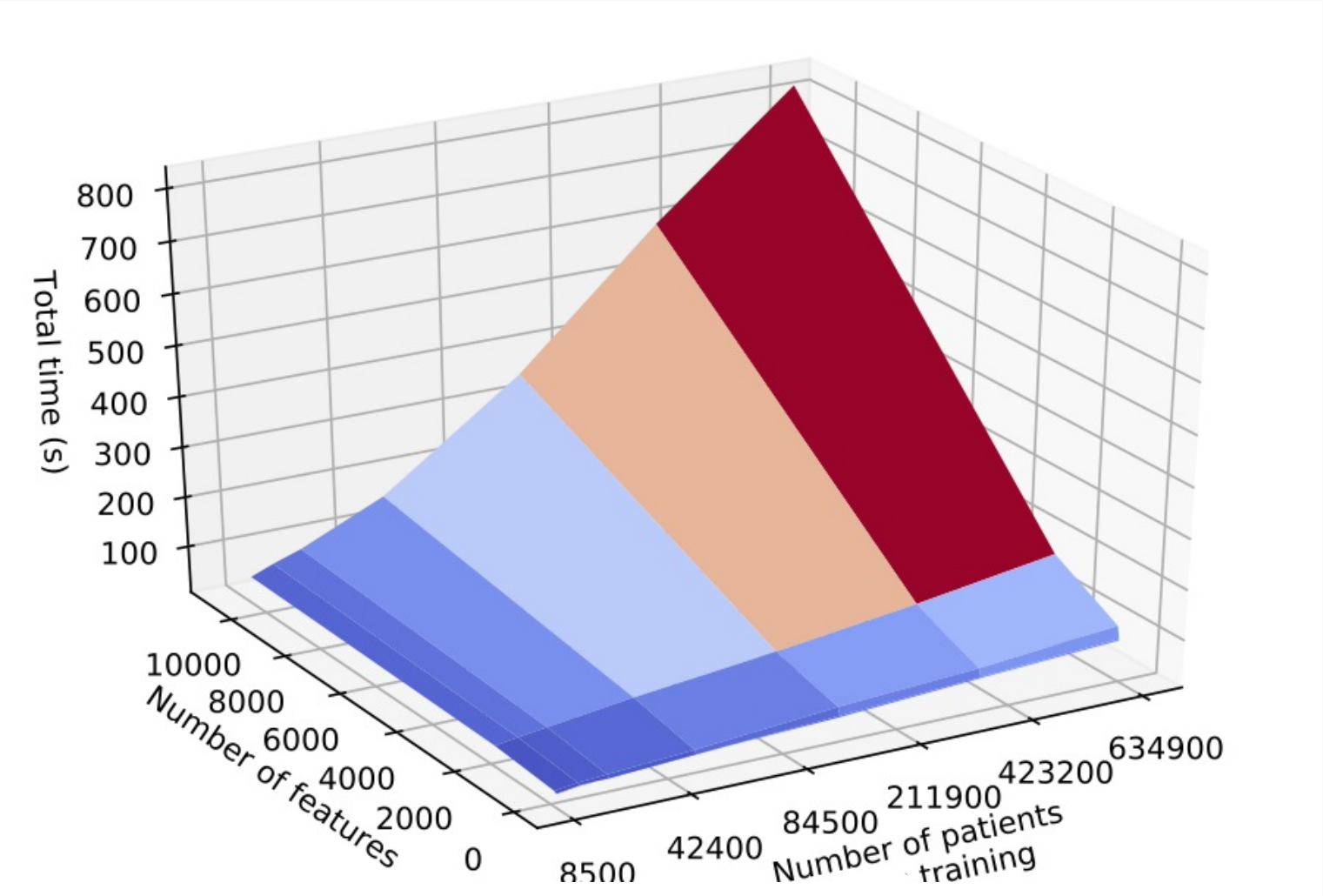
(c) ICU Admissions (AUC $\geq$ 65.6%).



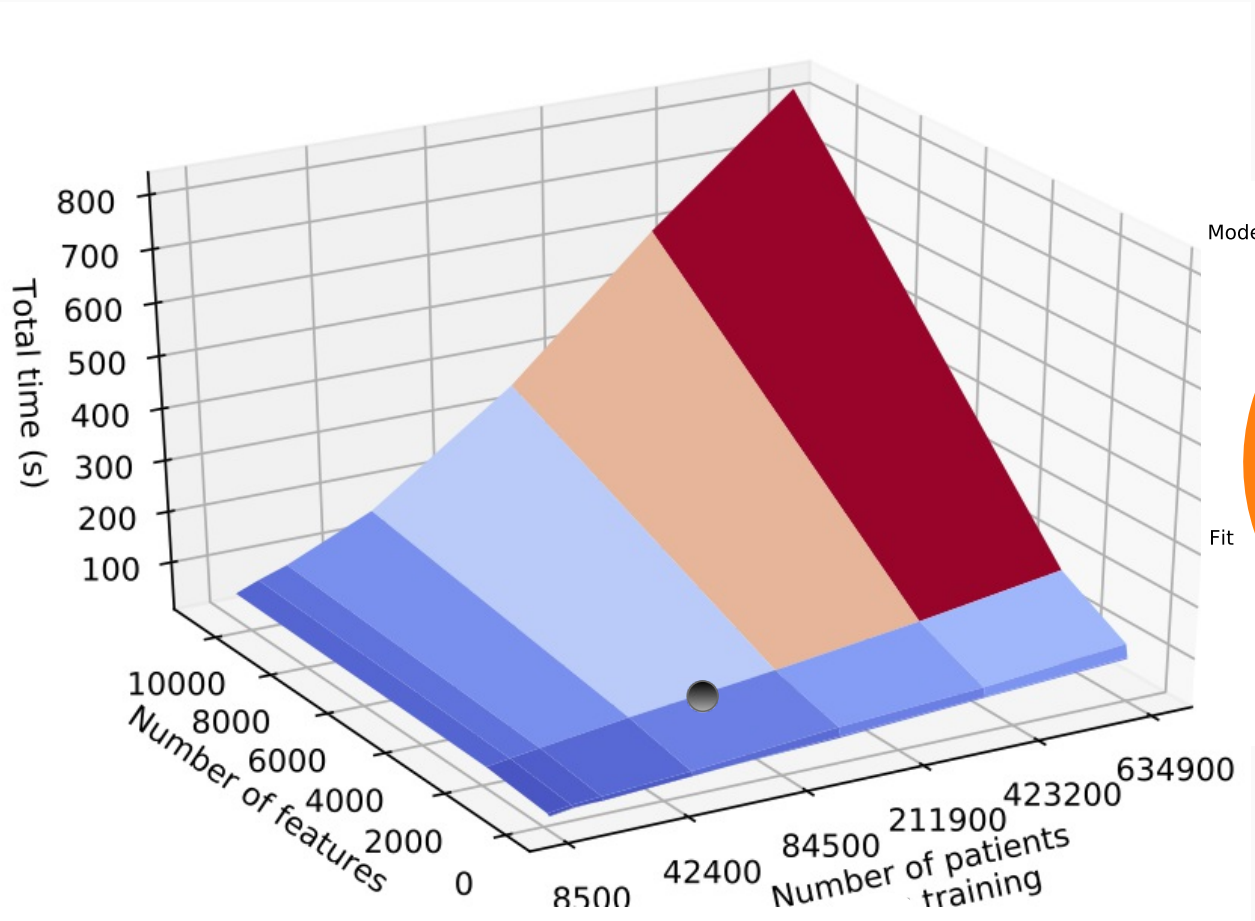
(d) Hospital-Acquired Infections (AUC $\geq$ 80%).

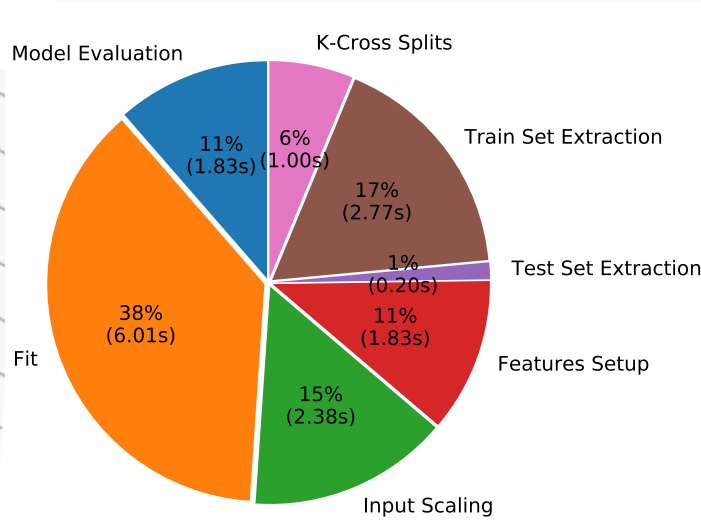
Figure 1: Predicting with **BMC** Features.

Total Computational Cost (LR 3-Cross)



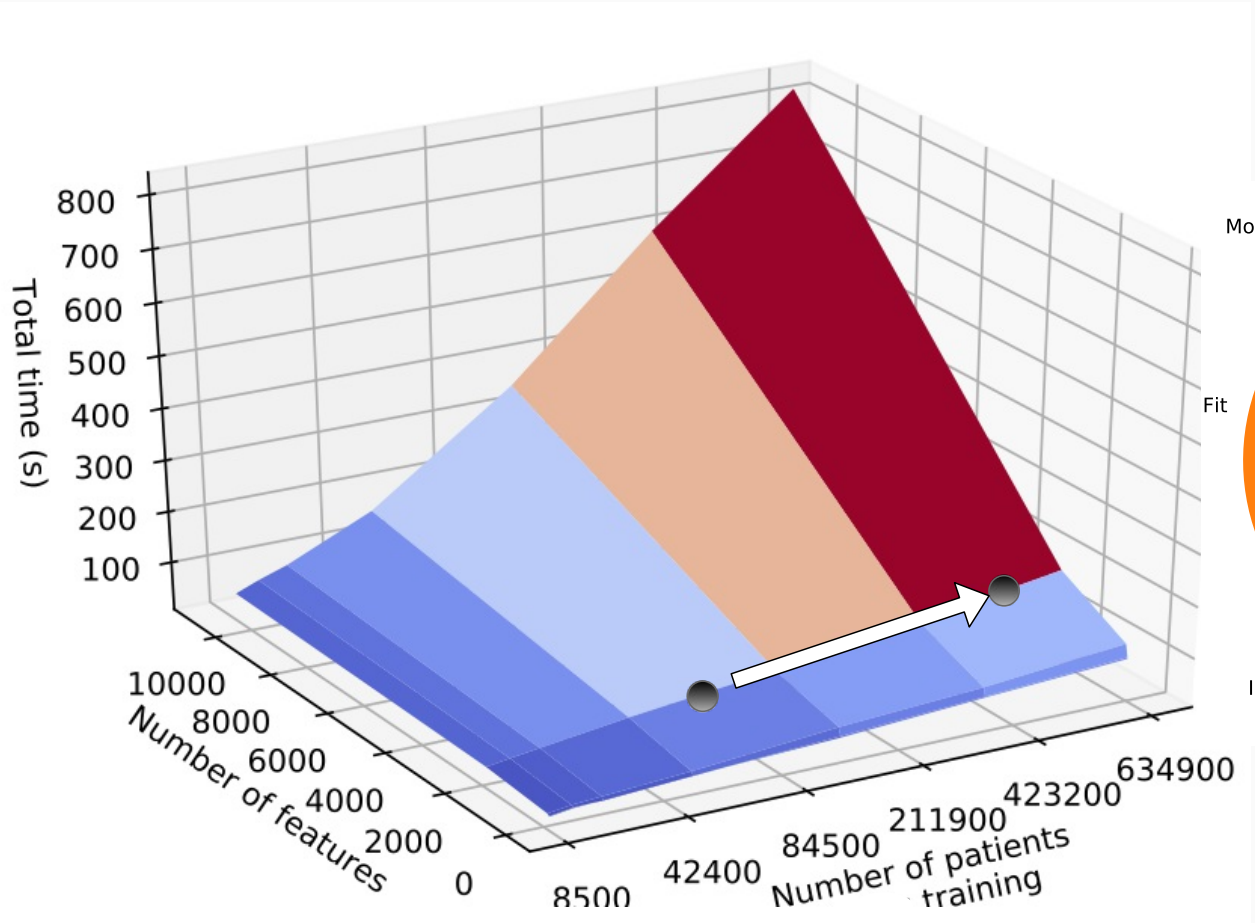
Zoom: Cost Breakdown

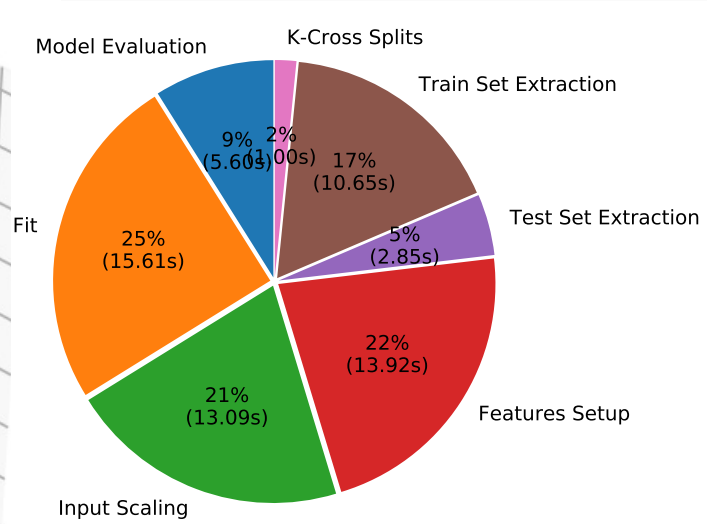




84 500 instances in train set, 2056 features

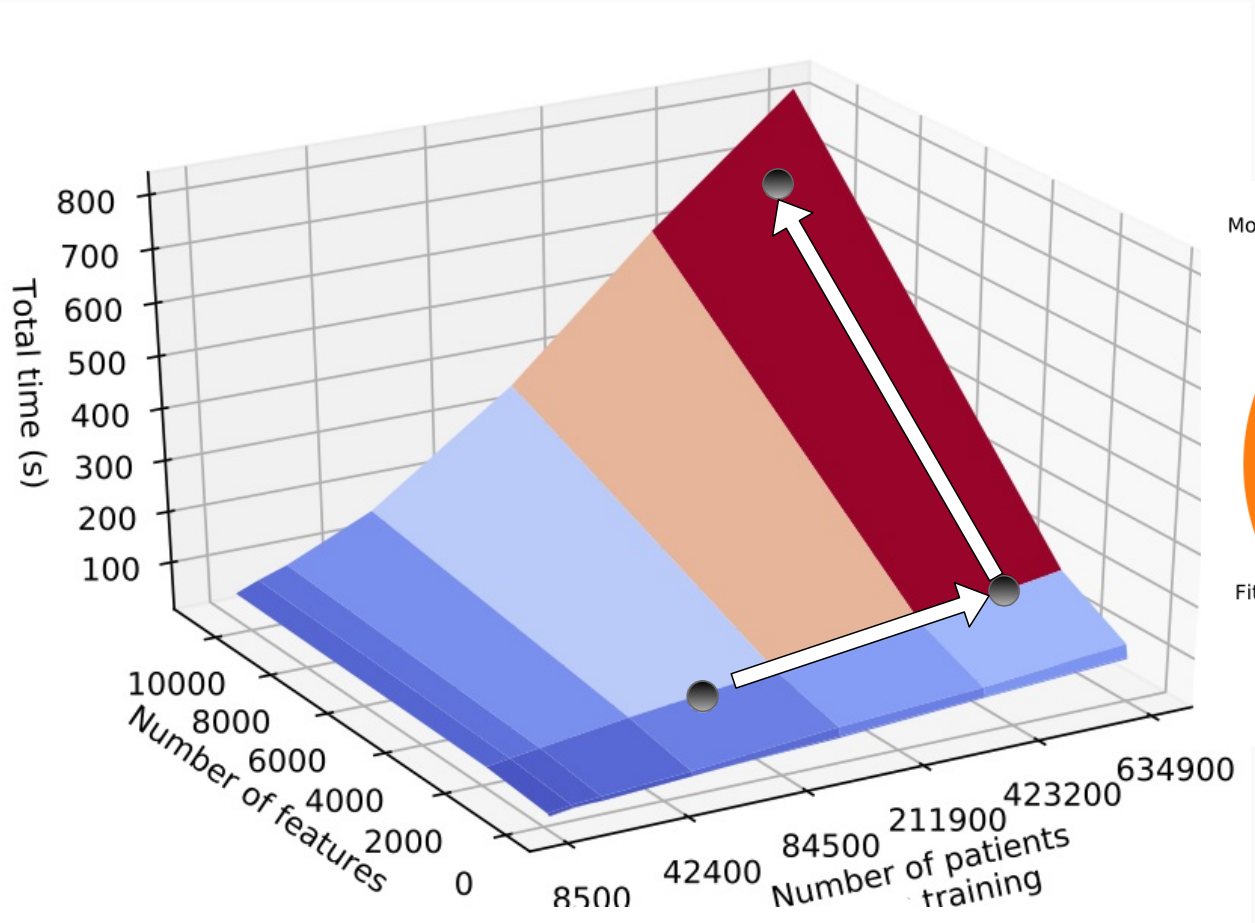
Zoom: Cost Breakdown

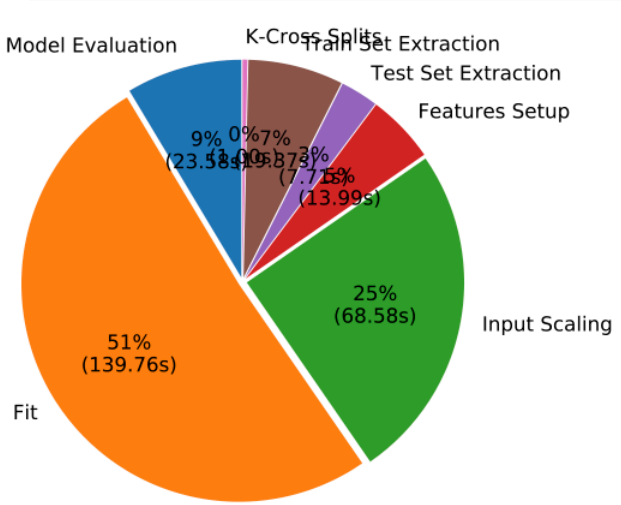




7x more instances in train set

Zoom: Cost Breakdown

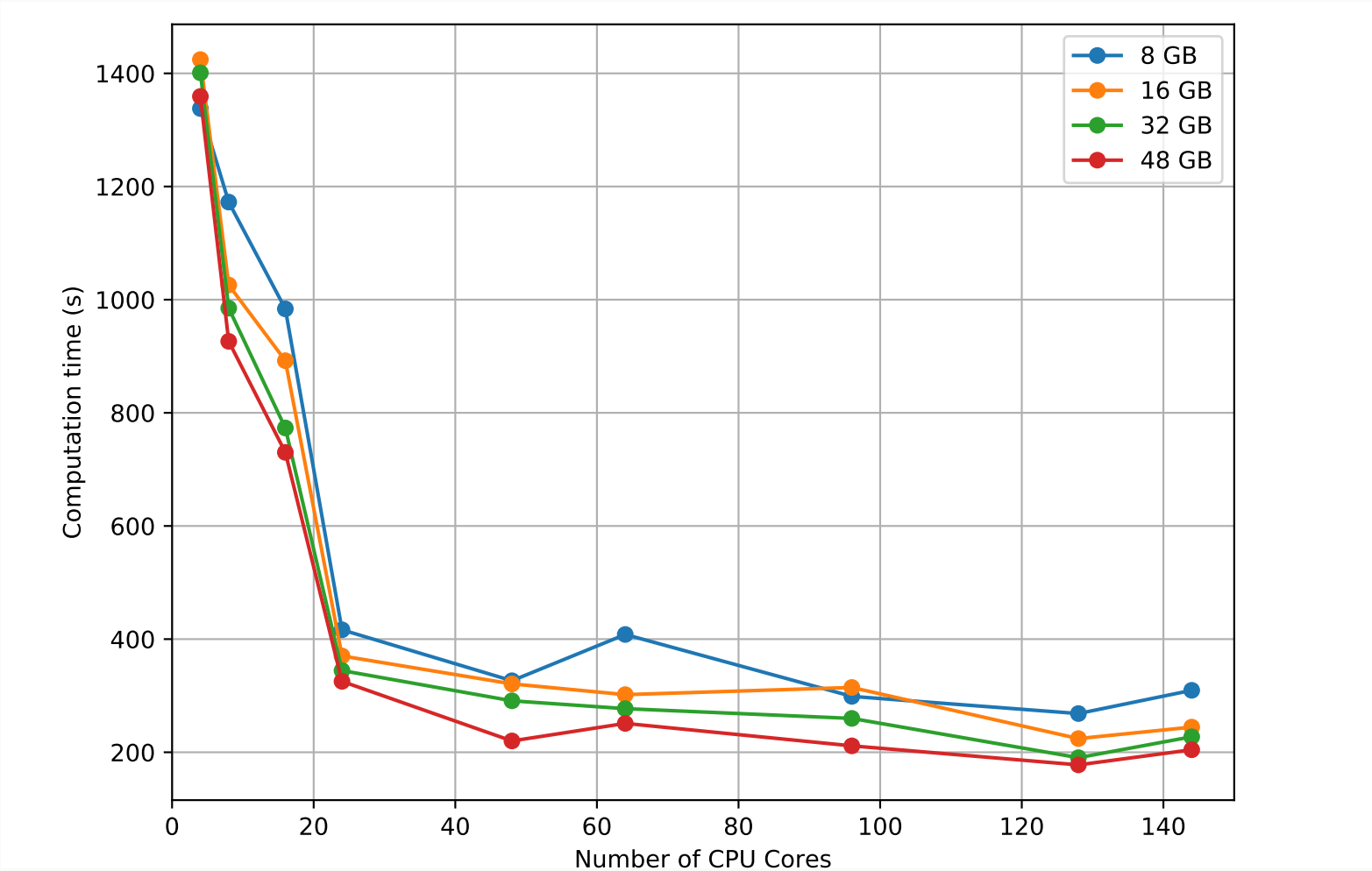




5x more features

Scalability with Hardware Resources

Total time for construction and 3-fold cross validation of BMP models w.r.t. number of cores and RAM·per·executor:



→ for this use case, it is possible to set resources to maintain an interactive session for the domain expert ($\geq$ 48 cores, $\geq$ 16 Gb RAM)

Scalability with Hardware Resources

- Performance with 48 and 144 cores almost similar (48Gb RAM)
- Increasing the number of cores (from 48 to 64, from 128 to 144) can worsen performance

What happens?

- Data partitioning matters
- more cores → more computational power, more partitions → more data shuffling (transfer)
- less cores → less partitions → less data transfert

Non-trivial balance to be found

Conclusions on the Use Case

1. Initial postulate reasonable: massive drug prescription data useful for prediction.

2. More variety (finer-grained features) $\rightarrow$ greater model accuracy

   (not systematic)

3. More volume $\rightarrow$ greater model accuracy

4. **Acceptable**[2] running times $\rightarrow$ good distribution of data & computations

[2] The domain expert ("clinical data scientist"?) can use **interactive sessions** for processing million of instances and thousands of features.

Article: Big Data Research, 2018

Scalable Machine Learning for Predicting At-Risk Profiles Upon Hospital Admission. Pierre Genevès, Thomas Calmant, Nabil Layaïda, Marion Lepelley, Svetlana Artemova, Jean-Luc Bosson.
Big Data Research, March 2018

Available from: http://tyrex.inria.fr/publications

Perspectives

Application-specific

- Clinical interpretation (LR models are intelligible, stable weights)
- More sophisticated features and models
- More data, e.g. drug molecular composition

Distributed data-centric programming

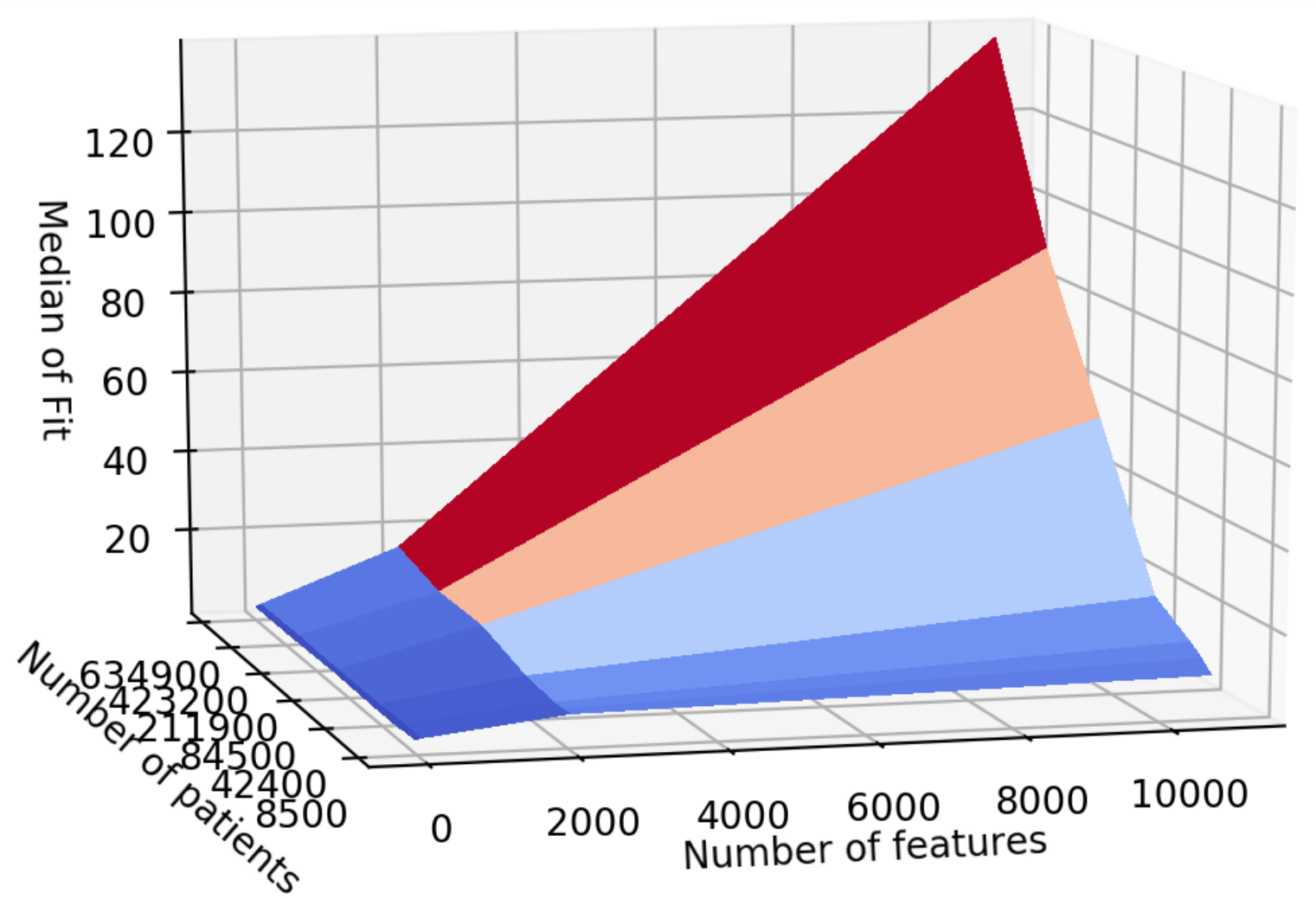
- Traditional complexity inappropriate (data transfer)
  - → Designing cost-models for these data-driven applications
- Scalability is not trivial to obtain
  - → optimizing compilers: synthesis of efficient distributed code

⋆ The CLEAR research project: http://tyrex.inria.fr/clear

Thank you!

Appendices

Computational Cost (LR model fit)



Labels

Considered Adverse Effects (AE):

(1) Death during hospital stay (3.00%: 44 667 cases)
(2) Admission to ICU on or after the second day (excluding patients directly admitted to ICU on the first day) (3.42%)
(3) Pressure ulcers not present at admission (2.55%)
(4) Hospital-acquired infections (2.54%)

Labels for the train set obtained automatically

- 1 boolean per AE, established from International Classification of Diseases (ICD9) codes in the database
- Labeling algorithm obtained with the help of clinicians

Classes are imbalanced

On 1 487 867 admissions, $> 8\%$ experience at least one AE $\in \{2, 3, 4\}$